Using bilingual word-embeddings for multilingual collocation extraction

Marcos Garcia, Marcos García-Salido and **Margarita Alonso-Ramos**
Universidade da Coruña, Departamento de Galego-Portugués, Francés e Lingüística
Facultade de Filoloxía, Campus da Zapateira, 15701 — A Coruña, Galicia, España
{marcos.garcia.gonzalez,marcos.garcias,margarita.alonso}@udc.gal

Abstract

This paper presents a new strategy for multilingual collocation extraction which takes advantage of parallel corpora to learn bilingual word-embeddings. Monolingual collocation candidates are retrieved using Universal Dependencies, while the distributional models are then applied to search for equivalents of the elements of each collocation in the target languages. The proposed method extracts not only collocation equivalents with direct translation between languages, but also other cases where the collocations in the two languages are not literal translations of each other. Several experiments —evaluating collocations with three syntactic patterns— in English, Spanish, and Portuguese show that our approach can effectively extract large pairs of bilingual equivalents with an average precision of about 90%. Moreover, preliminary results on comparable corpora suggest that the distributional models can be applied for identifying new bilingual collocations in different domains.

1 Introduction

Even though there is no universal definition of collocation, there is a general tendency to consider any syntactically related frequent pair of words to be a collocation (Smadja, 1993; Evert and Kermes, 2003; Kilgarriff, 2006). In the Firthian tradition of the term "collocation", not even a syntactic relation between the members is necessary, but in the phraseological tradition, not only the syntactic relation is a condition but also a lexical restriction.[1] From this phraseological point of view, a collocation is a restricted binary co-occurrence of lexical units (LUs) between which a syntactic relation holds, and that one of the LUs (the *base*) is chosen according to its meaning as an isolated LU, while the other (the *collocate*) is chosen depending on the base and the intended meaning of the co-occurrence as a whole, rather than on its meaning as an isolated LU (Mel'čuk, 1998). Thus, a noun in English such as "picture" requires the verb "to take" (and not "to do", or "to make") in the phrase "take a picture", while "statement" selects "to make" ("make a statement").

In a bilingual (or multilingual) scenario, equivalent collocations are needed to produce more natural utterances in the target language(s). In this regard, the referred noun "picture" would select the verb "tirar" in Portuguese —"to remove"— ("tirar uma fotografia"). Similarly the Spanish "vino" ("wine") would require the adjective "tinto" ("vino tinto"), which is not the main translation of "red" ("red wine").

The unpredictability of these structures involves problems for tasks such as machine translation, whose performance can benefit from lists of multilingual collocations (or transfer rules for these units) (Orliac and Dillinger, 2003). In areas like second language learning, it has been shown that even advanced learners need to know which word combinations are allowed in a specific linguistic variety (Altenberg and Granger, 2001; Alonso-Ramos et al., 2010). Thus, obtaining resources of multilingual equivalent collocations could be useful for different applications such as those mentioned above. However, this kind of resources is scarce, and constructing them manually requires a large effort of expert lexicographers.

In the last years, several approaches were implemented aimed at extracting bilingual collocations, both from parallel corpora (Kupiec, 1993;

[1] An overview of different visions on collocations —both from theoretical and practical perspectives— can be found in Seretan (2011).

Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), pages 21–30,
Valencia, Spain, April 4.

Smadja et al., 1996; Wu and Chang, 2003), and from comparable or even from non-related monolingual resources (Lü and Zhou, 2004; Rivera et al., 2013), often combining statistical approaches with the use of bilingual dictionaries to find equivalents of each *base*.

In this paper we explore the use of distributional semantics (by means of bilingual word-embeddings) for identifying bilingual equivalents of monolingual collocations: On one hand, monolingual collocation candidates are extracted using a harmonized syntactic annotation —provided by Universal Dependencies (UD)[2]—, as well as standard association measures. On the other hand, bilingual word-embeddings are trained using lemmatized versions of noisy parallel corpora. Finally, these bilingual models are employed to search for semantic equivalents of both the *base* and the *collocate* of each collocation.

Several experiments —using the OpenSubtitles2016 parallel corpora in English, Portuguese, and Spanish (Lison and Tiedemann, 2016)— show that the proposed method successfully identifies bilingual collocations with different patterns: *adjective-noun*, *noun-noun*, and *verb-object*. Furthermore, preliminary results in comparable corpora suggest that the same strategy can be applied in this kind of resources to extract new pairs of bilingual collocations.

Section 2 includes some related work on collocation extraction, specially on papers dealing with bilingual resources. Then, our method is presented and evaluated in Sections 3 and 4, respectively. Finally, some conclusions and further work are drawn in Section 5.

2 Related work

Several approaches were employed in order to automatically identify monolingual collocations (and other multiword expressions) from corpora. Most strategies use statistical association measures on windows of *n-grams* with different sizes (Church and Hanks, 1990; Smadja, 1993). Other methods, such as the one presented in Lin (1999), started to apply dependency parsing aimed at better identifying combinations of words which occur in actual syntactic relations. More recently, the large availability of better parsers allowed researchers to combine automatically obtained syntactic information with statistical methods to extract collocations more accurately (Evert, 2008; Seretan, 2011).

A different perspective on collocation extraction focuses not only on their retrieval, but on semantically classifying the obtained collocations, in order to make them more useful for NLP applications (Wanner et al., 2006; Wanner et al., 2016).

Concerning the extraction of bilingual collocations, most works rely on parallel corpora to find the equivalent of a collocation in a target language. In this respect, Smadja (1992; 1996) first identifies monolingual collocations in English (the source language), and then uses *mutual information* (MI) and the *Dice coefficient* (respectively) to find French equivalents of the source collocations.

Kupiec (1993) also uses parallel corpora to find noun phrase equivalents between English and French. The method consist in applying an expectation maximization (EM) algorithm to previously extracted monolingual collocations.

Similarly, Haruno et al. (1996) obtain Japanese-English chunk equivalents by computing their MI scores and taking into account their frequency and position in the aligned corpora.

Another work which uses parallel corpora is presented in Wu and Chang (2003). The authors extract Chinese and English *n-grams* from aligned sentences by computing their *log-likelihood* ratio. Then, the *competitive linking algorithm* is used to decide whether each bilingual pair actually corresponds to a translation equivalent.

More recently, Seretan and Wehrli (2007) took advantage of syntactic parsing to extract bilingual collocations from parallel corpora. The strategy consist in first extracting monolingual collocations using *log-likelihood*, and then searching for equivalents of each *base* using bilingual dictionaries. The method also uses the position of the collocation in the corpus, and relies on the syntactic analysis by assuming that equivalent collocations will occur in the same syntactic relation in both languages.

Rivera et al. (2013) present a framework for bilingual collocation retrieval which can be applied —with different modules— in parallel and in comparable corpora. As in other works, monolingual collocations (based on *n-grams*) are extracted in a first step, and then bilingual dictionaries (or WordNet, in the comparable corpora scenario) are used to find the equivalents of the *base* in the aligned sentence (or in a small window of

[2] http://universaldependencies.org/

adjacent sentences) of the source collocation.

A different approach, which uses non-related monolingual corpora for finding bilingual collocations, was presented in Lü and Zhou (2004). Here, the authors apply dependency parsing and the *log-likelihood* ratio for obtaining English and Chinese collocations. Then, they search for translations using word translation equivalents with the same dependency relation in the target language (using the EM algorithm and a bilingual dictionary).

Although not focused on collocations, Pascale Fung applied methods based on distributional semantics to build bilingual lexica from comparable corpora (Fung, 1998, among others). This approach takes into account that in this type of resources the position and the frequency of the source and target words are not comparable, and also that the translations of the source words might not exist in the target document.

Similarly, the approach presented in this paper leverages noisy parallel corpora for building bilingual word-embeddings. However, with a view to applying it in other scenarios (such as comparable corpora), it does not need information about the position of the collocations in the corpora, —neither their comparative frequency— to identify the equivalents. Furthermore, it does not take advantage of external resources such as bilingual dictionaries, so the method can be easily applied to other languages.

3 Bilingual collocation extraction

This section presents our method for automatically extracting bilingual collocations from corpora. First, we briefly describe the approach for identifying candidates of monolingual collocations using syntactic dependencies. Then, the process of creating the bilingual word-embeddings is shown, followed by the strategy for discovering the collocation equivalents between languages.

3.1 Monolingual dependency-based collocation extraction

Early works on *n-gram* based collocation extraction already pointed out the need for using syntactic analysis for better identifying collocations from corpora (Smadja, 1993; Lin, 1999). Syntactic analysis can, on the one hand, avoid the extraction of syntactically unrelated words which occur in a small context windows. On the other hand, it can effectively identify the syntactic relation between lexical items occurring in long-distance dependencies (Evert, 2008).

Besides, and even though it is not always the case (Lü and Zhou, 2004), our method assumes that most bilingual equivalent of collocations bear the same syntactic relation in both the source and the target languages.

In order to better capture the syntactic relations between the *base* and the *collocate* of each collocation, our method uses state-of-the-art dependency parsing. Apart from that, and aimed at obtaining harmonized syntactic information between languages, we rely on *universal dependencies* annotation, which permits the use of the same strategy for extracting and analyzing the collocations in multiple languages.[3]

Preprocessing: Before extracting the collocation candidates from each corpus, we apply a pipeline of NLP tools in order to annotate the text with the desired information. Thus, the output of this process consists of a parsed corpus in a CoNLL-U format, where for each word we have its surface form, its lemma, its POS-tag and morphosyntactic features, its syntactic head as well as the *universal* relation the word has in this context.[4]

From this analyzed corpus, we extract the word pairs belonging to the desired relations (collocation candidates). On the one hand, we keep their surface forms, POS-tags, and other syntactic dependents which may be useful for the identification of potential collocations. On the other hand, in order to apply association measures, we retain a list of triples containing (a) the syntactic relation, (b) the head, and (c) the dependent (using their lemmas together with the POS-tags). Thus, from a sentence such as "John took a great responsibility", we obtain (among others) the following triples:

$$\textit{nsubj}(\text{take}_{\text{VERB}},\text{John}_{\text{PROPN}})$$
$$\textit{amod}(\text{responsibility}_{\text{NOUN}},\text{great}_{\text{ADJ}})$$
$$\textit{dobj}(\text{take}_{\text{VERB}},\text{responsibility}_{\text{NOUN}})$$

This information (and also the corpus size and the frequency of the different elements of the potential collocations) is saved in order to rank the candidates.

Collocation patterns: At the moment, we are focused on extracting three different syntactic pat-

[3] http://universaldependencies.org/u/dep/all.html
[4] http://universaldependencies.org/format.html

terns of collocations in three languages (Spanish —*es*—, Portuguese —*pt*—, and English —*en*):

Adjective—Noun (amod): these candidates are pairs of adjectives (*collocate*) and nouns (*base*) where the former syntactically depends of the latter in a *amod* relation. Example: $killer_{base};serial_{collocate}$.

Noun—Noun (nmod): this collocation pattern consists of two common nouns related by the *nmod* relation, where the head is the *base* and the dependent is the *collocate* (optionally with a *case* marking dependent preposition: "of" in English, "de" in Portuguese and Spanish). Example: $rage_b;fit_c$.[5]

Verb—Object (vobj): *verb-object* collocations consists of a verb (the *collocate* and a common noun (the *base*) occurring in a *dobj* relation. Example: $care_b;take_c$.

Identification of candidates: For each of the three patterns of collocations, we extract a list of potential candidates for the three languages. After that, the candidates are ranked using standard association measures that have been widely used in collocation extraction (*MI, t-score, z-score, Dice, log-likelihood*, etc.) (Evert, 2008).

In the current experiments, we selected two statistical measures whose results complement each other: *t-score* (which prefers frequent dependency pairs, and has been proved useful for collocation extraction (Krenn and Evert, 2001)), and *mutual information* (which is useful for a large corpus (Pecina, 2010), even if it tends to assign high scores to candidates with very low-frequency).

The output of both association measures is merged in a final list for each language and collocation pattern, defining thresholds of *t-score*$=>2$ and *MI*$=>3$ (Stubbs, 1995), and extracting only collocations with a frequency of *f*$=>10$ (a relatively large value for reducing the extraction of incorrect entries from a noisy corpus and from potential errors of the automatic analysis).

It must be noted that, since these lists of monolingual collocations have been built based on statistical measures of collocability, their members need not be *bona fide* collocations in the phraseological meaning. Thus, the lists can include idioms (e.g., "kick the bucket"), quasi-idioms (e.g., "big deal") (Mel'čuk, 1998), or free combinations (e.g., "buy a drink").

[5]Note that some collocations belonging to this pattern are analyzed in UD —mainly in English— using the *compound* relation, so they are not extracted in the experiments performed in this paper.

3.2 Bilingual word-embeddings

Word-embeddings are low-dimensional vector representations of words which capture their distributional context in corpora. Even though distributional semantics methods have been largely used in previous years, approaches based on word-embeddings have gained in popularity recently, since the publication of *word2vec* (Mikolov et al., 2013).

Based on the *Skip-gram* model of *word2vec*, Luong et al. (2015) proposed *BiSkip*, a word-embeddings model which learns learns bilingual representations using aligned corpora, thus being able to predict words crosslinguistically.

As our approach for collocation extraction uses lemmas (instead of surface forms) to identify the candidates, the bilingual models are also trained on lemmatized corpora. Therefore, we convert the raw parallel corpora in lemmatized resources (with any other information) keeping the original sentence alignment.

Once we have the lemma version of the corpora, the bilingual models are built using MultiVec, an implementation of *word2vec* and *BiSkip* (Berard et al., 2016). As we work with three different languages, we need three different bilingual models: *es–en*, *es–pt*, and *pt–en*.

As it will be shown, the obtained models can predict the similarity between words in bilingual scenarios by computing the cosine distance between their vectors. As the models learn the distribution of single words (lemmas), they deal with different semantic phenomena such as polysemy or homonymy. Concerning collocations, this means that, ideally, the bilingual models could predict not only the equivalents of a *base*, but also to capture the (less close) semantic relation between the bilingual *collocates*, if they occur an enough number of times in the corpora.

3.3 Bilingual collocation alignment

In order to identify the bilingual equivalent (in a target language) of a collocation, our method needs (a) monolingual collocations (ideally obtained from similar resources), and (b) a bilingual *source-target* model of word-embeddings.

With these resources, the following strategy is applied: For each collocation in the source lan-

guage (e.g., *líob;tremendoc*, in Spanish) we select its *base* and obtain —using the bilingual model— the *n* most similar lemmas in the target language (where *n=5* in our experiments): "trouble", "mess", etc. Then, starting from the most similar lemma, we search in the target list for collocations containing the equivalents of the *base* (*troubleb;littlec*, *troubleb;deepc*, *messb;hugec*, *messb;finec*, etc.). If a collocation with a *base* equivalent is found, we compute the cosine distance between both *collocates* ("tremendo" *versus* "little", "deep", "huge", and "fine") and select them as potential candidates if their similarity is higher than a threshold (empirically defined in this paper as 0.65), and if the target candidate is among the *n* most similar words of the source *collocate* (again, *n=5*). Finally, if these conditions are met, we align the source and target collocations, assigning the average distance between the *bases* and the *collocates* as a confidence value: *es-en:líob;tremendoc=messb;hugec*;0.721.

4 Experiments

This section presents the experiments carried out in order to evaluate the proposed method (henceforth DIS) in the three analyzed languages, using the three collocation patterns defined in Section 3.1. Our approach is compared against a baseline system (BAS) which uses hand-crafted bilingual dictionaries.[6]

Corpora: Monolingual collocations were extracted from a subset of the OpenSubtitles2016 corpus (Lison and Tiedemann, 2016), which contains parallel corpora from TV and Movie subtitles. We selected this resource because it is a large and multilingual parallel corpus likely to contain different collocations types (also from an informal register) to those present in other corpora, thus being useful for comparative studies.[7]

From the *en*, *es* and *pt* corpora, we selected those sentences which appear in the three languages (a total of $13,017,016$). They were tokenized, lemmatized and POS-tagged with a multilingual NLP pipeline (Garcia and Gamallo, 2015), obtaining three corpora of $\approx 91M$ (*es* and *pt*), and $\approx 98M$ (*en*) tokens. The resulting data were enriched with syntactic annotation using statistical models trained with MaltParser (Nivre et al., 2007) and the 1.4 version of the UD treebanks (Nivre et al., 2016).

Lg	amod		nmod		vobj	
es	480k	13,870	1.6M	5,673	430k	17,723
pt	420k	12,967	1.7M	5,643	560k	20,984
en	460k	14,175	1.6M	3,133	490k	15,492

Table 1: Number of unique input dependencies for each syntactic pattern, and final monolingual collocation candidates.

Collocations: From each corpus, three patterns of collocations candidates were extracted: *amod*, *nmod*, and *vobj*. For each language and pattern, we obtained a single list of collocations by merging the *MI* and *t-score* outputs as explained in Section 3.1. Table 1 shows the number of filtered collocations in each case.

Another version of the corpora was created only with the lemma of each token, keeping the original sentence alignments. These corpora were used for training three bilingual word-embeddings models with MultiVec (with 100 dimensions and a window-size of 8 words): *es–en*, *es–pt*, and *pt–en*.[8]

Baseline (BAS): The performance of the method described in Section 3.3 was compared to a baseline which follows the same strategy, but using bilingual dictionaries instead of the word-embeddings models. Thus, the BAS method obtains the equivalents of both the *base* and the *collocate* of a source collocation, and verifies whether exists a target collocation with the translations. The bilingual dictionaries provided by the *apertium* project (SVN revision 75,477) were used for these experiments (Forcada et al., 2011).[9]

The *es-pt* dictionary has $14,364$ entries, while the *es-en* one contains $34,994$. The *pt-en* dictionary (not provided by *apertium*) was automatically obtained by transitivity from the two other lexica, with a size of $9,160$ pairs.

4.1 Results

With a view to knowing the performance of both BAS and DIS in the different scenarios, 100 bilingual collocation pairs were randomly selected

[6]The extractions of both methods are available at `http://www.grupolys.org/~marcos/pub/mwe17.tar.bz2`

[7]Note, however, that OpenSubtitles2016 includes non-professional translations with some noisy elements such as typos or case inconsistencies, among others.

[8]These models are available at `http://www.grupolys.org/~marcos/pub/mwe17_models.tar.bz2`

[9]`https://svn.code.sf.net/p/apertium/svn/`

Lg	amod		nmod		vobj	
Pair	BAS	DIS	BAS	DIS	BAS	DIS
es-pt	657	9,464	320	3,867	529	12,887
es-en	248	7,778	32	890	183	8,865
pt-en	213	7,083	43	917	241	9,206

Table 2: Number of bilingual extractions of the baseline and DIS systems.

from each language and pattern,[10] creating a total of 18 lists (9 from BAS and 9 from DIS).

Two reviewers labeled each bilingual collocation pair as (a) correct, (b) incorrect, or (c) dubious (which includes pairs where the translation might be correct in some contexts even if they were not considered faithful translations).[11] Correct collocation equivalents are those pairs where the monolingual extractions were considered correct (both in terms of co-occurrence frequency and of collocational pattern classification), and that their translations were judged as potential translations in a real scenario. The reviewers achieved 92% and 83% inter-annotator agreement in BAS and DIS outputs, respectively. Those pairs with correct/incorrect disagreement were discarded for the evaluation. Those with at least one dubious label were checked by a third annotator, deciding in each case whether they were correct, incorrect, o dubious.

From these data, we obtained the precision values for each case by dividing the number of correct collocation equivalents by the number of correct, incorrect, and dubious cases (so dubious cases were considered incorrect). Recall was obtained by multiplying the precision values for the number of extracted equivalents, and dividing the result by the smallest number of input collocations for each pair (Table 2).[12] Finally, we obtained F-score values (the harmonic mean between precision and recall) for each case, and calculated the macro-average results for each language, pattern, and approach.

Table 2 contains the bilingual collocation equivalents extracted by each method in the 9 settings, from the input lists of monolingual data (Table 1). These results clearly show that the baseline approach extract a lower number of bilingual equivalents. This might have happened due to the size of the dictionaries and because of the internal properties of the collocations, where the *collocates* may not be direct translations of each other. Moreover, it is worth noting that in both BAS and DIS results, the bilingual extractions including English are smaller than the *es-pt* ones.

Concerning the performance of the two approaches, Tables 3 (baseline) and 4 (DIS) contain the precision, recall and f-score for each language pair and collocation pattern.

BAS obtains high-precision results in every language and collocation pattern (91.7% in the worst scenario), with a macro-average value of 97%. These results are somehow expected due to the quality of the hand-crafted dictionaries. However, because of the poor recall numbers, the general performance of BAS is low, achieving f-score results of $\approx 4.7\%$. Interestingly, the size of the dictionary does not seem crucial to the results of the baseline. In this respect, the *es-pt* results are much higher (specially in recall) than *es-en*, whose dictionary size is more than the double. Also, the *pt-en* results are slightly better than the *es-pt* ones, the latter being obtained using a dictionary built by transitivity.

About DIS model, its precision is lower than the baseline, with results between 83.9% (*pt-en:vobj*) and 92.9% (*es-pt:amod*). However, this approach finds much more bilingual equivalents than the bilingual dictionaries, so recall values increase to an average of almost 50%. Unlike BAS (whose results are more homogeneous along the collocation patterns), DIS model obtains more variable numbers in each setting. Noticeably, the *nmod* extractions of the pairs including English have very low recall when compared to the other results, maybe derived from not having extracted nouns analyzed as *compound* (Section 3.1). As in the baseline, the DIS *es-pt* results are better than the two other pairs, so the linguistic distance seems to play an important role on bilingual collocation extraction.

The method proposed in this paper assigns a confidence value (obtained from the cosine distance between the vectors of the *base* and the *col-*

[10] Except for those baseline extractions with less than 100 elements, where all of them were selected.

[11] Some of these dubious equivalents are actual translations in the original corpus, such as the *es-en* "copa de champaña" ("champagne cup") — "cup of wine", even if they are semantically different.

[12] Note that these recall results assume that every collocation in the shortest input list of each pair has an equivalent on the other language, which is not always the case. Thus, more realistic recall values (which would need an evaluation of every extracted pair) will be higher than the obtained in our experiments.

Lang Pair	amod			nmod			vobj			avg		
	Prec	*Rec*	*F1*	*Prec*	*Rec*	*F1*	*Prec*	*Rec*	*F1*	*Prec*	*Rec*	*F1*
es–pt	99.0	5.0	9.6	97.8	5.5	10.5	98.7	3.0	5.7	98.5	4.5	8.6
es–en	95.8	1.7	3.4	100	1.0	2.0	100	1.2	2.3	98.6	1.3	2.6
pt–en	97.9	1.6	3.2	91.7	1.3	2.5	92.1	1.4	2.8	93.9	1.4	2.8
avg	97.6	2.8	5.4	96.5	2.6	5.1	96.9	1.9	3.6	97.0	1.8	4.7

Table 3: Precision, recall and f-score of the baseline (BAS) system (*avg* is macro-average).

Lang Pair	amod			nmod			vobj			avg		
	Prec	*Rec*	*F1*	*Prec*	*Rec*	*F1*	*Prec*	*Rec*	*F1*	*Prec*	*Rec*	*F1*
es–pt	92.9	67.8	78.4	93.8	64.3	76.3	90.1	66.0	76.5	92.5	66.0	77.1
es–en	92.0	51.6	64.3	88.0	25.0	38.9	84.0	48.1	61.2	87.5	41.6	56.4
pt–en	90.5	49.5	64.0	90.0	26.3	40.1	83.9	49.9	62.6	88.2	41.9	56.8
avg	91.8	56.3	68.9	90.6	38.5	51.9	86.2	54.7	66.7	89.5	49.8	63.4

Table 4: Precision, recall and f-score of DIS system (*avg* is macro-average).

locate equivalents) to each bilingual pair of collocations. In this respect, Figure 1 plots the average performance and confidence curves versus the total number of extracted pairs. This figure shows that using a high confidence value ($> 90\%$), it is possible to extract $\approx 35,000$ bilingual pairs with high-precision. Besides, it is worth mentioning that filtering the extraction with confidence values higher than 90% does not increase the precision of the system, so we can infer that the errors produced in the most confident pairs arise due factors other than the semantic similarity (e.g. different degrees of compositionality). However, as the confident value decreases the precision of the extraction also gets worse, despite the rise in the number of extractions which involves higher recall and consequently better f-score.

Finally, all the bilingual collocations extracted by DIS were merged into a single list with the three languages, thus obtaining new bilingual equivalents (not extracted directly by the system) by transitivity.[13] This final multilingual resource has $31,735$ collocations, $8,747$ of them with translations in the three languages.

4.2 Error analysis

The manually annotated lists of bilingual collocations were used to perform an initial error analysis of our approach. These errors were classified, due to its origin, in the following types:

[13]The merging process obtained $3,352$ new bilingual collocation equivalents not present in the original extractions.

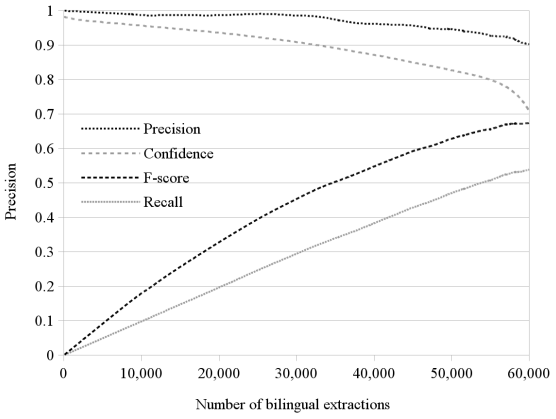

Figure 1: Average precision, recall, f-score, and confidence curves (from 0 to 1) versus total number of extractions of the DIS model.

Preprocessing: Several errors derived from issues produced by the NLP pipeline, such as POS-tagging or dependency parsing: e.g., "$\text{pain}_{\text{Noun}}$, end_{Verb}" was labeled as *dobj* (instead of *nsubj*).

Bilingual model: The bilingual word-embeddings approach, though useful, produces some errors such as the identification of antonyms (with similar distribution), which can align opposite collocation equivalents (such as *pt-en:tecido$_b$;vivo$_c$=tissue$_b$;dead$_c$*) where the extracted equivalent of the *collocate* "vivo" ("alive", in *pt*) was "dead". In most cases, however, the system obtained similar (but not synonym) collocations: *pt-en:chá$_b$;preto$_c$=coffee$_b$;black$_c$* ("black tea, black coffee").

Lemmatization and gender: The lemmatization of some words differs from language to language, so working with lemmas instead of tokens also might involve some errors. For instance, the word "hija" ("daughter", in Spanish) is lemmatized as "hijo" ("son") in Spanish and Portuguese ("filha, filho"), while in English "son" and "daughter" appear as different entries. Thus, some bilingual collocations differ in the gender of their *bases*: *es-en:hijo$_b$;encantador$_c$=daughter$_b$;lovely$_c$*

Monolingual extraction: The extraction of *base* and *collocate* pairs produced incorrect collocations such as *plan$_b$;figure$_c$*, instead of obtaining the phrasal verb "figure out" as *collocate*.

Other errors: Some other errors were produced by mixed languages in the original corpus (e.g., the verb form "are", in English, was analyzed as a verb form of the verb "arar" —"to plow"—, in Spanish) and from noise and misspellings in the corpora (proper nouns with lower case letters, etc.).

4.3 Comparable corpora

A final experiment was carried out in order to know (a) whether the bilingual word-embeddings —trained in the same parallel corpora as those used for extracting the collocations— could be successfully applied for aligning collocations obtained from different resources, and (b) the performance of the proposed method in comparable corpora.

So we applied the same strategy for monolingual collocation extraction in the Spanish and Portuguese *Wikipedia Comparable Corpus 2014*,[14] and calculated the semantic similarity between the collocations using the same word-embeddings models as in the previous experiments.

From these corpora, we obtained filtered lists of $73,291$ and $119,311$ candidate collocations in Portuguese and Spanish, respectively (from 140M, and 80M of tokens). From the $51,183$ bilingual collocations obtained by the DIS approach, we randomly selected and evaluated 100 *es-pt* pairs.

The precision of the extraction was 88.9%, with a recall of 62.1% (again computed using the whole set of monolingual collocations), and 73.1% f-score. These results are in line with those obtained in the OpenSubtitles *es-pt* pair ($\approx 3\%$ lower), so the method works well in different corpora and domains. It is worth noting that $43,025$ of the extracted collocation equivalents (84%) had not been retrieved from the OpenSubtitles corpus.

This last experiment shows that (a) the bilingual word-embeddings can be used for identifying collocation equivalents in different corpora than those used for training, and that (b) they can also be applied in corpora of different domains to obtain previously unseen multilingual collocations.

5 Conclusions and further work

In this paper we have presented a new strategy to automatically discover multilingual collocation equivalents from corpora.

First, three different patterns of monolingual collocations were extracted using syntactic analysis provided by harmonized UD annotation, together with a combination of standard association measures.

Besides, bilingual word-embeddings were trained in parallel corpora that had been previously lemmatized. These bilingual models were then used to find distributional equivalents of both the *base* and the *collocate* of each source collocation in the target language.

The performed experiments, using noisy parallel corpora in three languages, showed that the proposed method achieves an average precision in the bilingual alignment of collocations of about 90%, with reasonable recall values. Furthermore, the evaluation pointed out that using a confidence value for setting up a threshold is useful for retaining only high-precise bilingual equivalents, which could benefit different work on multilingual lexicography.

Finally, a preliminary test using comparable corpora suggested that the bilingual word-embeddings can be efficiently applied in different corpora than those used for learning, discovering new bilingual collocations not present in the original resources.

In further work, the results of the error analysis should be taken into account in order to reduce both the errors produced by the NLP pipeline, and those which arise from the word-embedding models. In this respect, it could be interesting to evaluate other approaches for the alignment of bilingual collocations which make use of better compositionality models and which effectively learn the semantic distribution of collocations.

[14] http://linguatools.org/tools/corpora/wikipedia-comparable-corpora/

Acknowledgments

This work has been supported by the Spanish Ministry of Economy, Industry and Competitiveness (MINECO) trough the projects with reference FFI2016-78299-P and FFI2014-51978-C2-1-R, by a *Juan de la Cierva formación* grant (FJCI-2014-22853), and by a postdoctoral fellowship granted by the Galician Government (POS-A/2013/191).

References

Margarita Alonso-Ramos, Leo Wanner, Orsolya Vincze, Gerard Casamayor del Bosque, Nancy Vázquez Veiga, Estela Mosqueira Suárez, and Sabela Prieto González. 2010. Towards a Motivated Annotation Schema of Collocation Errors in Learner Corpora. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010)*, pages 3209–3214, Paris. European Language Resources Association (ELRA).

Bengt Altenberg and Sylviane Granger. 2001. The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied linguistics*, 22(2):173–195.

Alexandre Berard, Christophe Servan, Olivier Pietquin, and Laurent Besacier. 2016. MultiVec: a Multilingual and Multilevel Representation Learning Toolkit for NLP. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4188–4192, Paris. European Language Resources Association (ELRA).

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Stefan Evert and Hannah Kermes. 2003. Experiments on candidate data for collocation extraction. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics (EACL 2003)*, volume 2, pages 83–86, Budapest. Association for Computational Linguistics.

Stefan Evert. 2008. Corpora and collocations. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook*, volume 2, pages 1212–1248. Mouton de Gruyter, Berlin.

Mikel L. Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Pascale Fung. 1998. A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas. Machine Translation and the Information Soup (AMTA 1998)*, pages 1–17, Langhorne, Pennsylvania. Association for Machine Translation in the Americas.

Marcos Garcia and Pablo Gamallo. 2015. Yet Another Suite of Multilingual NLP Tools. In José-Luis Sierra-Rodríguez and José Paulo Leal and Alberto Simões, editor, *Languages, Applications and Technologies. Communications in Computer and Information Science*, International Symposium on Languages, Applications and Technologies (SLATE 2015), pages 65–75.

Masahiko Haruno, Satoru Ikehara, and Takefumi Yamazaki. 1996. Learning bilingual collocations by word-level sorting. In *Proceedings of the 16th Conference on Computational Linguistics (COLING 1996)*, volume 1, pages 525–530, Copenhagen. Association for Computational Linguistics.

Adam Kilgarriff. 2006. Collocationality (and how to measure it). In Elisa Corino and Carla Marello and Cristina Onesti, editor, *Proceedings of the 12th EURALEX International Congress*, volume 2, pages 997–1004, Torino.

Brigitte Krenn and Stefan Evert. 2001. Can we do better than frequency? A case study on extracting PP-verb collocations. In *Proceedings of the ACL Workshop on Collocations*, pages 39–46, Toulouse. Association for Computational Linguistics.

Julian Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics (ACL 1993)*, pages 17–22, Columbus, Ohio. Association for Computational Linguistics.

Dekang Lin. 1999. Automatic Identification of Non-compositional Phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics (ACL 1999)*, pages 317–324, College Park, Maryland. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente

Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris. European Language Resources Association (ELRA).

Yajuan Lü and Ming Zhou. 2004. Collocation translation acquisition using monolingual corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL 2004)*, pages 167–174, Barcelona. Association for Computational Linguistics.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual Word Representations with Monolingual Quality in Mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing (VSM-NLP) at the 2015 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT 2015)*, pages 151–159, Denver, Colorado. Association for Computational Linguistics.

Igor Mel'čuk. 1998. Collocations and Lexical Functions. In Anthony Paul Cowie, editor, *Phraseology. Theory, Analysis and Applications*, pages 23–53. Clarendon Press, Oxford.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Workshop Proceedings of the International Conference on Learning Representations (ICLR) 2013*, Scottsdale, Arizona. arXiv preprint arXiv:1301.3781.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(02):95–135.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666, Paris. European Language Resources Association (ELRA).

Brigitte Orliac and Mike Dillinger. 2003. Collocation extraction for machine translation. In *Proceedings of Ninth Machine Translation Summit (MT Summit IX)*, pages 292–298, New Orleans, Lousiana.

Pavel Pecina. 2010. Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44(1-2):137–158.

Oscar Mendoza Rivera, Ruslan Mitkov, and Gloria Corpas Pastor. 2013. A Flexible Framework for Collocation Retrieval and Translation from Parallel and Comparable Corpora. In *Proceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technology*, pages 18–25, Nice.

Violeta Seretan and Eric Wehrli. 2007. Collocation translation based on sentence alignment and parsing. In *Actes de la 14e conference sur le Traitement Automatique des Langues Naturelles (TALN 2007)*, pages 401–410, Toulouse.

Violeta Seretan. 2011. *Syntax-based collocation extraction*, volume 44 of *Text, Speech and Language Technology Series*. Springer Science & Business Media.

Frank Smadja, Kathleen R McKeown, and Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational linguistics*, 22(1):1–38.

Frank Smadja. 1992. How to compile a bilingual collocational lexicon automatically. In *Proceedings of the AAAI Workshop on Statistically-Based NLP Techniques*, pages 57–63, San Jose, CA.

Frank Smadja. 1993. Retrieving Collocations from Text: Xtract. *Computational linguistics*, 19(1):143–177.

Michael Stubbs. 1995. Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of language*, 2(1):23–55.

Leo Wanner, Bernd Bohnet, and Mark Giereth. 2006. Making sense of collocations. *Computer Speech & Language*, 20(4):609–624.

Leo Wanner, Gabriela Ferraro, and Pol Moreno. 2016. Towards Distributional Semantics-Based Classification of Collocations for Collocation Dictionaries. *International Journal of Lexicography*. 10.1093/ijl/ecw002.

Chien-Cheng Wu and Jason S Chang. 2003. Bilingual Collocation Extraction Based on Syntactic and Statistical Analyses. In *Proceedings of the 15th Conference on Computational Linguistics and Speech Processing (ROCLING 2003)*, pages 1–20, Taiwan. Association for Computational Linguistics and Chinese Language Processing.